**Project 4 – Group 6 Proposal**

Group:

Caleb Steeves

DJ Thapa

Tanner Victorian

Steffi Yang

Jessica Warner

Summary:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Dataset:

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

For the Machine Learning requirement, what column in your data will be target/labels/dependent variable? Is this a classification or regression problem?

* Our Target variable is labeled “stroke” and has a 1 if the patient had a stroke and a 0 if not.
* Our other variables consist of but not limited to; Gender, age, hypertension, heart disease, marital status, work type, residence type, bmi, smoking status.
* This would be considered a classification problem, with machine learning identifying based on our X variables impact predicting y variable (0 or 1).

What metric will you use to evaluate your model?

* We will use classification accuracy to evaluate our model.

What tool(s) will you use to transform/prepare your data?

* We will import our dataset into jupyter notebook/ Colab.
* Pyspark/ SQL → Clean up the data → convert to Pandas
* Use pandas to hot-in-code/ pandas dummies categorical variables
* Create a final data frame for Dependent and Target Variable

What tool(s) will you use to visualize your results?

* We will use pandas and Pyspark DataFrames; Matplotlib at the minimum

How will you divide the work?

1. Find data set (CSV)

Completed

1. Create spark session and convert to PySpark DF- DJ Completed
2. Clean and filter in spark- Tanner (Null values and removing the column id)
3. SQL+ Visualizations
   1. Create queries- DJ
   2. Visualizations (graphs)
   3. Regression Model- Variables impacting Stroke
4. Convert to Pandas DF - Jessica
   1. Encode categorical data into numerical → Either Hot-encode or Pandas get-dummies
   2. Create X and Y Variables
5. Create Machine learning models